UTEC-CSc-77-017
Semi-Annual Technical Report
Computer Science
December 1977

SENSORY INFORMATION PROCESSING

UNIVERSITY OF UTAH

Sponsored by
Defense Advanced Research Projects Agency
ARPA Order Number 2477

(6) SENSORY INFORMATION PROCESSING.

1 July 1975 THROUGH 31 December 1975

(9) Semi-Annual Technical Report.
1 Jul-31 Dec 75,

Contractor: University of Utah
Contract Number: (15) DAHC15-73-C-0363, ARPA Order-2477
Effective Date: 1 July 1973
Expiration Date: 31 January 1976
Amount of Contract: $2,512,000.00
Project Code: 3D30

(10)
Principal Investigator: Dr. Thomas G. Stockham, Jr.
Telephone: (801) 581-8224

(11) 31 Dec 75

Contracting Officer: Mr. Edgar S. Allen
DSSW

(12) 58 P.

Approved for public release;
distribution unlimited.

Sponsored by
Defense Advanced Research Projects Agency
ARPA Order Number 2477

(14) UTEC-CSc-77-017

40H949 JOB

TABLE OF CONTENTS

REPORT SUMMARY

In Section 1. Baxter discusses and demonstrates what parameter adjustments can accomplish in his method of removing motion blur from photographic images.

Section 2 indicates the status of work by Faugeras in color vision, and our expected method of reporting these results.

In Section 3. Boll describes a method for strikingly increasing the perceived quality of synthetic speech. Additional computation at the receiver is used to generate two channels (i.e. binaural) of sound for a stereo headphone set. This method requires no change in the existing generation and transmission processes and algorithms.

In Section 4. Petersen demonstrates one method of removing noise from speech. Intelligible speech has been generated from input signals that contain +18DB of noise.

In Section 5. Callahan reviews his method of suppressing noise in a one-dimensional signal stream (e.g. speech) by using two-dimensional processing. A complete Technical Report UTEC-CSc-76-209 will be available approximately concurrently with this report.

Sections 6. and 7. report the start of work in the coding of speech, and in the mathematical theory of human perception. Both indicate our future directions in these fields.

Section 8. outlines the Image Understanding Research by Newell that has been proposed for the future.

SECTION 1

REMOVING MOTION BLUR FROM PHOTOGRAPHIC IMAGES

Brent Baxter

In the previous semi-annual report a new method was described for removing blur from photographic images. Several experiments are described here illustrating the effect of adjusting the cutoff frequency of filters used in the restoration scheme.

The method operates by high pass filtering the logarithm of the blurred image spectrum and then adding the low pass log spectrum of a different, unblurred image. This combination accomplishes two important results: First, blurring represents a partial loss of signal energy at high spatial frequencies, and adding the prototype information after filtering tends to restore energy at these frequencies to the proper level. Second, excessive amplification of film grain is prevented. Signal energy at spatial frequencies near zeroes of the blur spectrum cannot be recovered completely and any attempt to do so will result in excessive amplification of film grain and other kinds of system noise. To see how this is avoided, note that zeroes in the blur spectrum become sharp, spike-like negative

impulses when the logarithm is taken and these impulses are preserved in the high pass filter operation, This prevents undesirable amplification at those spatial frequencies dominated by noise.

In the restored images described below, the low pass and high pass filters were constrained to have frequency responses, the sum of which was a constant.

$$LPF(\omega) = 1(\omega) - HPF(\omega)$$

This was done to avoid problems in preserving the average brightness of the image.

Adjustments to the cutoff frequency of these filters demonstrated that those filters with a basically circular shape tended to allow some of the predominant features of the prototype to appear in the restored image. This effect is shown in Figure 1. Tailoring the shape of the filter frequency response as described above tends to minimize the effect as shown in Figure 2. Figure 3 is the image used to construct the prototype spectrum. Notice the strong diagonal character in Figure 1 due to the ladder in the prototype image. In a practical system this effect could be minimized by averaging log spectra from several images as well as by using the noncircular frequency responses mentioned above. Using elongated filters has the effect of restoring the image only in the direction of the blur while

leaving it undisturbed in other directions.

Reducing the cutoff frequency of the filters tends to make the restoraton less noisy and also somewhat less sharp. Figures 4 and 5 are examples of restorations obtained using filter cutoff frequencies one eighth as high as those shown previously. At present there is no systematic way to select the optimum cutoff frequency, however, it may be varied over wide limits with only a moderate effect on the restoration.

The image chosen for this study (Figure 6) was taken with a pocket instamatic camera on 16 mm Kodacolor film, in a completely unrehearsed manner. The film is quite small and rather grainy resulting in a somewhat noisy restoration. Considerably better results are obtainable if care is taken to record the blurred image on a larger size, fine grain film. Fine grain development also helps. Nevertheless, the model seems to describe the blur process well enough for the method to work quite well. Stripes can be seen in the woman's blouse that were not at all visible in the original blurred image; their existence has been confirmed by comparison with the actual garment.

Correcting the phase of the image transform may be accomplished in a variety of ways as described by Cannon [1], but in the present case advantage was taken of a streak in the background formed by the reflection of a flash bulb in a spherical globe. The length and direction of the

streak was used to compute the phase of the blur and also to determine the proper orientation of the high pass filter. The presence of this streak made it possible to evaluate the effect of changing filter cutoff frequencies independently of the problem of estimating the phase and direction of the blur.

Periodic convolution was used for the filtering operations rather than the more common aperiodic techniques Cole [2] making it necessary to preprocess the borders of the image. The edges of the blurred image were simply extended out from the borders and smoothly scaled down to zero intensity. This operation is intended to simulate the effect of blurring across boundaries of adjacent copies of the image. It might be argued that this is not a good simulation of a blur, however, it works well enough; essentially no edge effects are noticeable in the restored image. See McDonnell and Bates [3] for additional discussion of this topic.

This method is capable of removing a wide class of blurs from ordinary photographic images when the exact nature of the blur system is not known. We are presently investigating the applicability of these ideas to more sophisticated blur removal processes such as the speckle interferometry technique of Laberyie [4].

FIGURES FOR SECTION 1

Figure 1 -- Image restored using filters having circular symmetry

Figure 2 -- Image restored using filters with elongated frequency responses oriented perpendicular to the direction of the blur

Figure 3 -- Sharp image used to create a prototype log spectrum

Figure 4 -- Image restored as in Figure 1 but with lower cutoff frequency filters

Figure 5 -- Image restored as in Figure 2 but with lower cutoff frequency filters

Figure 6 -- Blurred image taken with a pocket instamatic camera

FIGURE 1

Image restored using filters
having circular symmetry

FIGURE 2

Image restored using filt. s with elongated
frequency responses oriented perpendicular to
the direction of the blur

FIGURE 3

Sharp image used to create
a prototype log spectrum

FIGURE 4

Image restored as in Figure 1 but with
lower cutoff frequency filters

FIGURE 5

Image restored as in Figure 2 but with lower cutoff frequency filters

FIGURE 6

Blurred image taken with a pocket instamatic camera

REFERENCES

[1] T.M Cannon, "Digital Image Deblurring by Nonlinear Homomorphic Filtering" ARPA Technical Report UTEC-CSc-74-091

[2] E.R. Cole, "The Removal of Unknown Blurs by Homomorphic Filtering" ARPA Technical Report UTEC-CSc-74-029

[3] M.J. McDonnell and R.H.T. Bates, "Restoring Parts of Scenes from Blurred Photographs" Optics Communication Vol. 13, 3 (March 1975): pp 347-349

[4] A. Labeyrie, "Attainment of Diffraction Limited Resolution in Large Telescopes by Fourier Analysing Speckle Patterns in Star Images" Astron. Astrophys. Vol. 6 (1970); pp. 85-87.

SECTION 2

COLOR IMAGE PROCESSING

Oliver Faugeras

The theoretical background of the work of Faugeras was reported in the last Semi-annual. During this six months, the Comtal display has been integrated into our available computing facility, and experiments are continuing. Color photographic pictures have been produced, but the expected results are not yet available in a fully optimal manner. We anticipate that this work will be ready for publication in a separate Technical Report (TR) during the late summer of 1976. We expect to publish this TR with color pictures in a limited quantity, but we do not expect to include color pictures in the Semi-annual Report series.

SECTION 3

SPEECH ENHANCEMENT AND CODING IMPROVING SYNTHETIC SPEECH QUALITY USING BINAURAL REVERBERATION

Steven F. Boll

The degrading characteristics of synthetic speech such as minimum phase effects, pitch and voicing errors and spectral distortions are more evident when the speech is listened to on headphones than when heard in a room over a loudspeaker. Listening to monaural sound in a room over a loudspeaker differs from headphone listening in two major respects: one, a different sound source is presented to each (binaural reproduction); and two, the sound source is altered by the room's acoustics, (reverberation). An experiment was conducted to include the effects of binaural reverberation on synthetic speech heard on headphones. To achieve this effect, the impulse response of a 20' x 20' classroom was first measured by applying an electrical pulse to a loudspeaker and recording the resulting room-loudspeaker impulse response as measured by two microphones spaced the ears distance apart. Figure 1 shows the microphone placements within a dummy head. Figures 2 and 3 show the measured left and right impulse responses and Figures 4 and 5 the respective frequency responses. These impulse responses were then convolved with the speech and played through each headset channel. Results demonstrate

[1] that this process not only suppresses the characteristics distortions of the synthetic speech, but also externalizes the sound source giving the effect of non-headphone listening. An example showing how this process reduces the minimum-phase "buzzy" quality is given in Figures 6 through 8. Figure 6 is of an original vowel. Figure 7 is the corresponding synthetic speech with its abnormally high peak factor, and Figure 8 the post-processed, convolved response. An example showing how this process reduces the effect of a voicing error is given in Figures 9 through 11. Figure 9 is a segment of an original fricated /sh/, Figure 10 is the corresponding synthetic speech generated with the incorrect voicing decision, and Figure 11 the post-processed convolved response.

Matching and Coding of Nonlinear Spectral Estimates by Linear Prediction

Introduction. This research considered applying the spectral matching properties of linear prediction analysis to nonlinear spectral estimates. Three areas are considered: one, the matching and coding of the log magnitude spectrum by LPC (Cepstral Prediction); two, the estimation of spectral zeros as well as poles by matching to the derivative of the log magnitude spectrum, (ramp modulated cepstral prediction); and three, the modeling and

Figure 1

coding of the spectrum modified by middle and inner ear nonlinearities, (the inner spectrum).

Linear Predictive Spectral Matching. Linear Prediction defines a technique for matching a given power spectrum $\hat{P}(\omega)$ by an all-pole power spectrum $P(\omega)$. From its set of autocorrelations $R(k)$, a set of predictor coefficients, $a(k)$ and gain factor G are computed which minimize the loss function:

$$E = \frac{G^2}{2\pi} \int_{-\pi}^{\pi} \frac{P(\omega)}{\hat{P}(\omega)} \, d\omega$$

where

$$\hat{P}(\omega) = \frac{G^2}{\left| \sum_{k=0}^{p} a(k) e^{-jk\omega} \right|^2}$$

The coefficients $a(k)$ are computed by solving the toeplitz system of equations:

$$\sum_{k=1}^{p} a(k) \, R(i-k) = -R(i) \quad 1 \leq i \leq p$$

where

$$R(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} P(\omega) \cos(k\omega) d\omega$$

with

$$G^2 = R(0) + \sum_{k=1}^{p} a(k)R(k)$$

Thus the analysis procedure consists of starting with a given $P(\omega)$, inverse transforming to a set of R(k), and solving for a(k) and G

Nonlinear Spectral Matching. The same analysis procedure can be followed where now a function $f(P(\omega))$ replaces $P(\omega)$ in the matching process. The loss function to be minimized is given by

$$E' = \frac{H^2}{2\pi} \int_{\pi}^{\pi} \frac{f(P(\omega))}{\hat{T}(\omega)} d\omega$$

$\hat{T}(\omega)$ is an all-pole power spectrum which is computed to minimize E'. It is estimated using the same procedure as above, namely let

$$U(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(P(\omega)) \cos(k\omega) d\omega$$

and solve

$$\sum_{k=1}^{p} q(k)U(k-i) = -U(i) \quad 1 \leq i \leq p$$

for g(k) giving

$$\hat{T}(\omega) = \frac{H^2}{\left| \sum_{k=0}^{p} q(k)\bar{e}^{jk\omega} \right|^2}$$

with

$$H^2 = U(0) + \sum_{k=1}^{p} q(k)U(k)$$

Three areas of this approach of nonlinear spectral modeling are considered.

Cepstral Prediction. By matching to the $\log^2 P(\omega)$ and LPC estimate of the real cepstrum is obtained. This minimum phase cepstral estimate is uniquely defined by either its set of reflection or predictor coefficients. This method allows for waveforms to be analyzed by homomorphic techniques but coded using LPC parameters; and thus allows for the interfacing of a homomorphic vocoder analyzer with a transmission channel set up to transmit LPC parameters.

Ramp Modulated Cepstral Prediction. A linear prediction spectral match $\frac{d}{d\omega} \log P(\omega)$ is obtained by applying linear prediction analysis to the ramp modulated real cepstrum. The spectral peaks of the resulting filter are then matched to peaks representing both poles and zeros of the prototype spectrum. For prototype spectra where zero approximation is important (such as for matching narrow stop bands for dynamically changing noise suppression filters), the ramp modulated cepstrum analysis offers an efficient technique for zero estimation.

Matching Spectra Modified by Ear Nonlinearities. There is evidence that the spectrum of the signal applied to the external ear is modified by nonlinear distortion in the middle and inner ear. The resulting "inner spectrum" is the one converted by hair cells into neural discharges. Inclusion of approximations to these spectral nonlinearities prior to linear prediction matching is continuing to be

considered in order to improve speech quality of a low bandwith speech analysis-synthesis system.

A-D-A Diagnostic Checkout Procedures. In cooperation with the ARPA-IPT Network Speech Compression group, Utah provided a package of various programs and essential peripheral equipment specifications required in order to test and maintain A-D and D-A Converters, (NSC Note 74).

In the enclosed package were copies of the various programs and specifications of equipment used at Utah to maintain our 15 bit A-D converters and our 16 bit D-A converters. This package is intended to augment the procedures and theory developed in Chin Moh Tasi's Master Thesis entitled "A Digital Technique for Testing A-D and D-A Converters". Copies of Tsai's thesis were mailed to the NSC group on June 20, 1974. Included are the following:

1. Specifications for a low noise/distortion Krohn-Hite oscillator: Model 4024.
2. Specifications for an external distortion measurement system: Model 1700A (Sound Technology).
3. Descriptions of two programs used to test A-D and D-A converters: ADTEST.SAV and AUDIO.SAV.
4. A copy of "A-D and D-A Converter Testing and Maintenance/Photolog".

The theory supporting the checkout procedures is presented in Tsai's thesis. The programs ADTEST.SAV and AUDIO.SAV are two extended versions of the procedures defined in the thesis. They both make essentially the same measurements and thus only extended documentation of the latter is provided.

Waveform Matching Using Linear Predictive Coding.

Essential to the design of systems for speaker authentication, variable vocoder frame rate transmission, word spotting in continuous speech, and isolated word recognition, is the requirement for comparing a reference waveform to an input waveform. A theoretical development for comparing two waveforms using linear predictive coding was considered during the period 1 January 1975 through 30 June 1975. The results of this investigation were in University of Utah Computer Science Memorandum No. 7500, May 1975 and distributed to the ARPA Network Speech Compression group as NSC Note 60.

Based upon this theoretical development, an isolation word recognition system using the linear prediction residual was developed, [2], [3]. Results demonstrating the effectiveness of this method for comparing speech waveform were evidenced by a recognition accuracy of 98.1% when the vocabulary consisted of 107 flight commands having an average of two syllables per word.

Optimal Time Registration Using Dynamic Programming. Essential to any procedure used to compare two waveforms is the need for aligning the reference pattern with the input pattern. For this isolated word recognition program, a modified Dynamic Programming procedure was used. This technique defined a nonlinear time warping function which attempts to align two waveforms of different lengths so as to minimize the total distance between them.

The theoretical development and FORTRAN procedures required to implement the isolated word recognition system and Dynamic Programming time warping function were provided to the ARPA Network Speech Compression group as NSC Note 73.

REFERENCES

[1] S.F. Boll, E. Ferretti, T. Peterson, "Improving Synthetic Speech Quality Using Binaural Reverberation", Proc. of the IEEE Conf. on Acoustics, Speech and Signal Processing, Philadelpia, P.A., April 1976.

[2] M.J. Coker, "An Isolated-Word Recognition System Based On Linear Prediction Analysis," M.S. Thesis, Elec. Eng. Dept., Univ. of Utah, 1975.

[3] M. Coker and S.F. Boll, "An Improved Isolation Word Recognition System Based Upon the Linear Prediction Residual", Proc. of the IEEE Conf. on Acoustics, Speech and Signal Processing. Phildelphia, P.A. April 1976.

SECTION 4

NOISE SUPPRESSION WITH LINEAR PREDICTION FILTERING.

Tracy L. Petersen

The preceding semi-annual report[2] describes the development of a dynamic noise suppression filter based on linear prediction spectral modeling[1], where the formula for a Wiener filter is implemented to construct successive filter estimates over short time intervals. These estimates then determine the time-varying characteristics of a linear prediction lattice filter.

Continued work with this noise suppression model during this half year has focused on suppressing noise from noisy speech rather than singing voice. A series of experiments were conducted to determine the relationship between parameter conditions in the dynamic Wiener filter model and performance of the model in effectively suppressing noise from noisy speech. Main results were three-fold. The time increment between successive filter estimates was reduced from 150 milliseconds to 12 milliseconds which prevented excessive noise from appearing during rapid transitions in the speech. Maximum attenuation was increased from -24dB to -48dB, and filter k-parameters were reduced from 90 to 64. Following these modifications, tests were made where the

perceived suppression of noise was judged by experienced listeners to be in the neighborhood of 18dB while the filtered speech remained perfectly intelligible. Upper limits on the level of noise which may be successfully suppressed from noisy speech have not yet been determined, but results indicate that even higher levels of noise may be suppressed from noisy signals.

IMPROVING SYNTHETIC SPEECH QUALITY BY MODELING THE LPC DRIVING FUNCTION.

When an LPC synthesizer is excited with the speech error signal (the true pitch information) the speech is reconstructed as the original (within finite word length limits). When the error signal is coded as a pulse train for voiced speech or as white noise for unvoiced speech the resulting synthesized speech contains an undesireable distortion usually described as a "buzzy" quality. If the error signal could be simulated at the synthesizer directly from coded pitch information, presumeably synthetic speech quality would be greatly improved without increasing the bandwidth of the channel signals. Recent work by Petersen has focused on this problem specifically. It has been found that a strong correspondence exists between the structure of k-parameters (filter coefficients) derived from analysis on an error signal and pitch information coded from that error signal, allowing for the possibility of extrapolating a

time-varying set of filter coefficients from a prototype set based on the input pitch information. Some initial studies have shown that such a model, when driven with standard pitch pulses, produces a waveform with characteristics similar to the error signal. Furtner work and testing with this model will be required to determine its degree of usefulness in improving synthetic speech quality.

REFERENCES

[1] J. Makhoul, "Linear Prediction: A Tutorial Review", Proc. IEEE, Vol. 63, No.4, April 1975.

[2] T.L. Petersen, "Removal of Noise from an Audio Signal", Section 7 UTEC-CSc-76-008, June 1975.

SECTION 5

SPEECH PROCESSING TO REDUCE NOISE AND IMPROVE INTELLIGIBILITY

Michael Wayne Callahan

A new method of acoustic signal processing has been investigated which is based on the short-time spectrum, a two-dimensional representation that shows the frequency content of the signal as a function of time. This representation is appropriate for signals such as speech and music, where the natural frequencies of the source change. In addition, this representation similar to frequency analysis in the human auditory system, so that signal modifications can be related to perceptual criteria. This method has been applied successfully to removal of broadband background noise (signal-to-noise ratio about 30dB) and to removal of high level interfering signals with strong harmonic structure (signal-to-noise ratio about 26dB). Both of these experiments were described in previous reports.

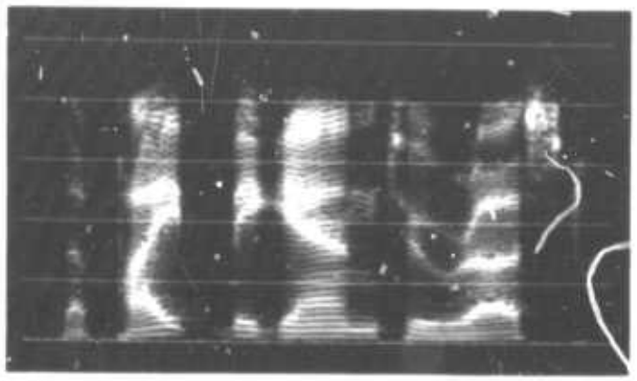
Research during the last period has been directed at isolating speech features in the short-time spectrum which are known to be important to perception, and to applying these results in a compression/expansion system for transmitting speech through a noisy channel.

Isolation Of Perceptually Important Speech Features:

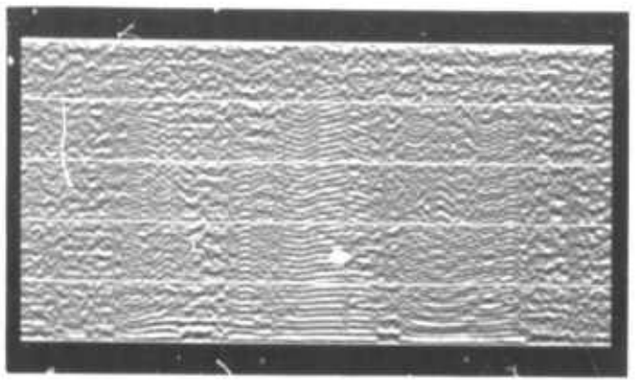
Experiments were conducted to attempt to isolate selected speech features by bandpass filtering the logarithm of the short-time spectrum. (The logarithm was taken to model approximate logarithmic sensitivity of the ear.) Three speech features were selected as typical: pitch, formants, and plosive noise bursts.

The results of these experiments are shown in Figures 1.b, 1.c, and 1.d, together with the original speech in Figure 1.a. The pictures have been scaled to use the full range of the film, so the figures do not show amplitude relative to the original speech. The pitch, formants, and plosive bursts have much lower dynamic range than the original. Most of the dynamic range of the original is in the slowly changing component of the short-time spectrum. This is illustrated by Figure 1.e, which shows features obtained by filtering the logarithm of the short-time spectrum in two dimensions to suppress the slowly changing component. The fact that the perceptually important features are still apparent in Figure 1.e suggests that speech processed in this manner should still be highly intelligible, and this is in fact the case. Such speech might therefore be more intelligible than normal speech in a noisy environment. Informal listening tests and the results discussed below support this notion.

(a)



(b)



(c)

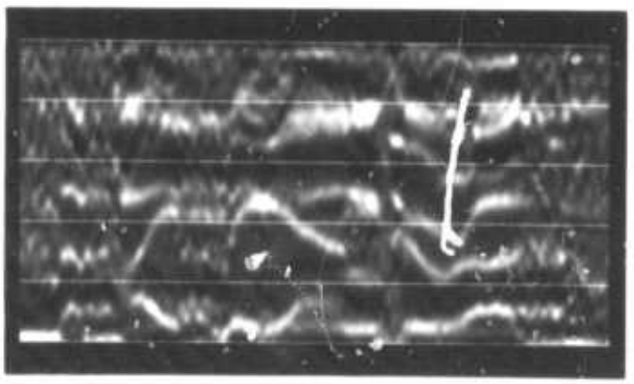
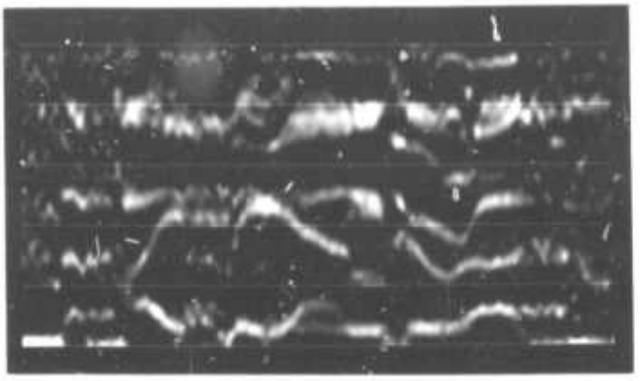


FIGURE 1

Speech features obtained by t) dimensional filtering of the logarithm of the short-time spectrum; (a) original speech, (b) pitch, (c) formants, (d) plosive noise bursts, (e) slowly changing component removed. The speech is "the pipe began to rust".

(continued)

(d)



(e)

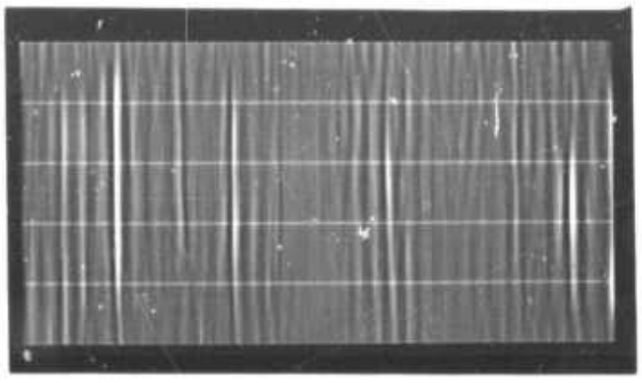


FIGURE 1

Speech features obtained by two dimensional filtering of the logarithm of the short-time spectrum; (a) original speech, (b) pitch, (c) formants, (d) plosive noise bursts, (e) slowly changing component removed. The speech is "the pipe began to rust".

Two-Dimensional Compression And Expansion Of Speech:

Oppenheim et al [1] investigated a homomorphic system for compression and expansion of acoustic signals. The system models audio signals as a product of two components - a slowly varying, positive envelope and a rapidly varying bipolar signal. These multiplied signals can be mapped into added signals by the logarithm and the envelope compressed or expanded by linear filtering. A two-dimensional (frequency/time) system for compression and expansion can be obtained in a similar manner by modeling the short-time spectrum as a product of two components - a slowly varying envelope which is of lesser importance, and a rapidly varying component which contains most perceptually important features. Compressing the signal by attenuating the large, slowly changing component should greatly reduce the dynamic range of the signal while preserving the information content.

A compression/expansion system based on this concept is shown in Figure 2, "STFT" represents the short-time Fourier transform, and "$\sum$" represents reconstruction of a time signal. The block "T" represents the effect of channel noise: e.g., tape hiss or quantization noise.

The two-dimensional system was simulated for both an analog and digital channel. In both cases the system

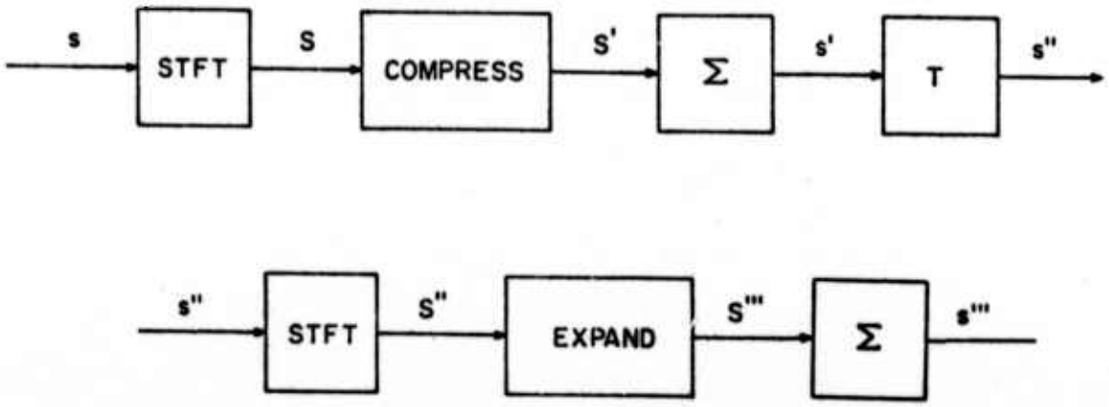


FIGURE 2

Two dimensional compression/expansion system for transmitting acoustic signals through a noisy channel.

provided considerable improvement over a similar homomorphic system. In the analog case, noise was first audible in the output of the two-dimensional system at a channel signal-to-noise ratio of 12dB, compared to 30dB for the one-dimensional homomorphic system. In the digital channel simulation, the two-dimensional system provided a three-bit improvement over the homomorphic system. Noise was first audible in the output of the two-dimensional system with four bit channel quantization, compared to seven bit quantization for the homomorphic system. As a reference, noise is audible in the uncompressed signal with 9 - 10 bit quantization.

Although distortions are audible, the system still produces natural, intelligible speech with three bit channel quantization.

REFERENCES

[1] Openheim, A.V., Shafer, R.W., T.G. Stockham, Jr. "Nonlinear Filtering of Multiplied and Convolved Signals", Proc. IEEE, Vol. 56, pp 1264-1291, August, 1968

SECTION 6

LINEAR PREDICTIVE CODING WITH A GLOTTAL WAVEFORM MODEL

William J. Done

Although synthetic speech generated with all-pole LPC techniques is intelligible and natural sounding, it suffers from a raspy or coarse sound. This flaw is especially annoying when the listener is exposed to long segments of synthetic speech. A possible source of this degradation is the failure of all-pole linear predictive models to accurately match zeros in the speech spectra or the failure to duplicate (using impulse excitation) the excitation characteristics of the actual glottal pulse wave.

The linear prediction technique is based on the approximation of the n^{th} speech sample s(n) as a summation of the N previous, linearly weighted samples to form $\hat{s}(n)$. An all-pole inverse filter is obtained by approximating the spectrum of the error as a constant. By relating the energy in the error sequence to the integral of the ratio of the speech power spectrum to the estimate's power spectrum, it is evident that the all-pole approximation will result in a good fit to the poles in the spectrum. By this same process, it is apparent that spectral zeros will not be matched as well.

To better model the zeros, it is assumed that the error signal

$$e(n) = s(n) - \hat{s}(n)$$

$$= s(n) - \sum_{i=1}^{M} a(i)s(n-i)$$

could be modelled as the effect of the glottal pulse on the zeros of the vocal tract. That is

$$s(n) = \sum_{i=1}^{N} a(i)s(n-i) + e(n)$$

$$= \sum_{i=1}^{N} a(i)s(n-i) + \sum_{j=0}^{M} b(j)g(n-j) \ ,$$

where the b(j) are the zero coefficients and g(n) is an assumed glottal waveform model. Note that the zero coefficients are not excited during times when g(n) = 0. This corresponds to a closed glottis condition. The analysis-synthesis procedure based on this model for voiced speech is summarized as follows:

1. The pole coefficients, a(i), are determined by

linear predictive analysis using the covariance technique.

2. e(n) is generated from the original speech and the pole coefficients.
3. Using the least squares method, the zero coefficients, $\tilde{b}(j)$, are generated from the assumed glottal waveform g(n) and the error signal e(n).
4. Synthesis is performed by using e(n) as the excitation,

$$s(n) = \sum_{i=1}^{N} a(i)s(n-i) + \tilde{e}(n)$$

where e(n) is generated from the glottal waveform model and the zero coefficients calculated in step 3:

$$\tilde{e}(n) = \sum_{j=o}^{M} \tilde{b}(j)g(n-j)$$

For unvoiced speech the all-pole model excited by noise is used [1].

The previous discussion proposes a glottal waveform model to develop the zero coefficients. The models used are

based on results by Rosenberg [2] and Holmes [3]. Construction of the glottal wave for each segment of voiced speech is based on the pitch value for that segment. The waveform is generated with a fixed amplitude and pitch dependent opening and closing times. Figure 1 illustrates the two models used to date. In the figure, T_p is the pitch period, T_0 the opening time, and T_C the closing time. The smooth model of Figure 1a) is constructed of polynomic segments. This was the first model tried, based on naturalness tests reported in [2]. When the smooth model failed to reproduce the rapid transitions in the error signal, the triangular pulse of Figure 1b) was developed. It achieved better results in duplicating the sharp transitions of the error signal.

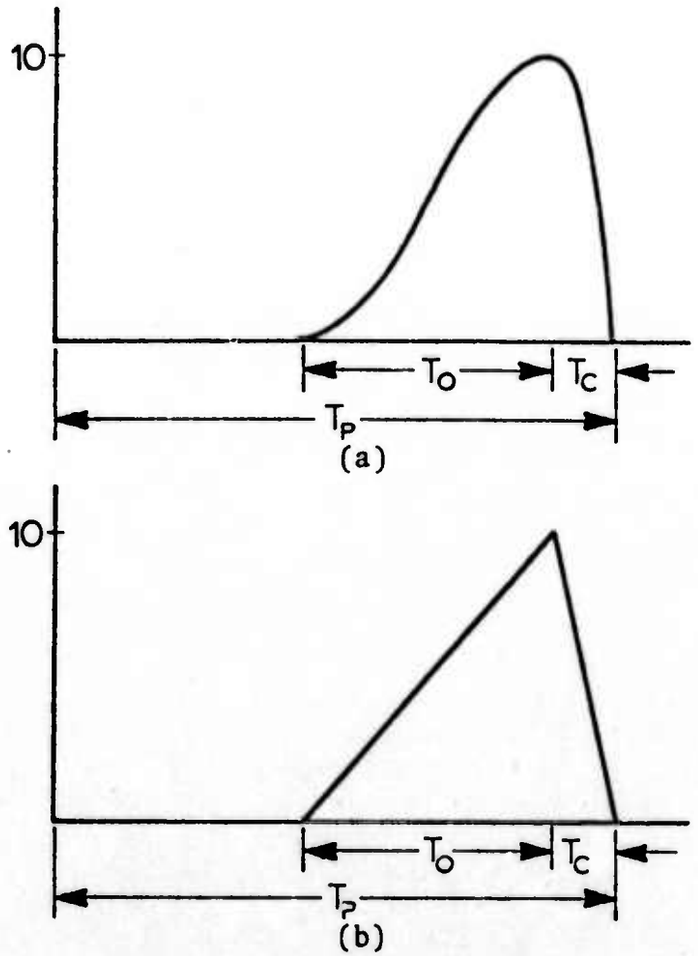


Figure 1. Glottal Waveform Models.

In analyzing the error signal, it is apparent that the peak-to-peak excursions of the error waveform are maximum in the neighborhood of an open glottis. For these segments, the glottal pulse is forcing the vocal tract, and the speech waveform begins to deviate from the response due to vocal tract characteristics. The sudden closure of the glottis corresponds approximately to the large excursions in the next pitch period. Because of the changing waveform characteristics strongly evident in these areas, the error signal resulting from a linear predictive model grows in magnitude during these intervals. In order to approximate these segments more closely than intervals when the glottis is closed, the error signal is zeroed for the intervals approximately corresponding to a closed glottis, as determined from the original speech waveform. Figure 2 illustrates this process.

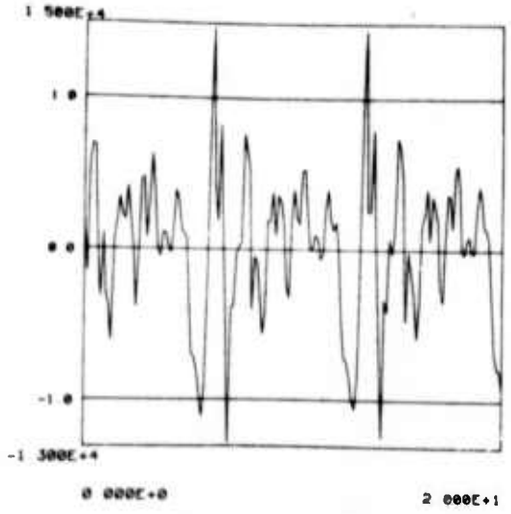
As mentioned, the triangular glottal pulse model was needed to better approximate the sharp transitions of the error signal. M, the order of the zero model, and the opening and closing times of the glottal pulse influence the match of the approximate error signal to e(n), the original error waveform. For example, glottal parameter settings of $T_0 = 0.24T_p$ and $T_C = 0.04T_p$, and an order of M = 6 (7 zero coefficients) produce a good representation of the error signal when the triangular model is used. Figure 2d) gives the approximate error signal resulting from this modelling

for the vowel /æ/.

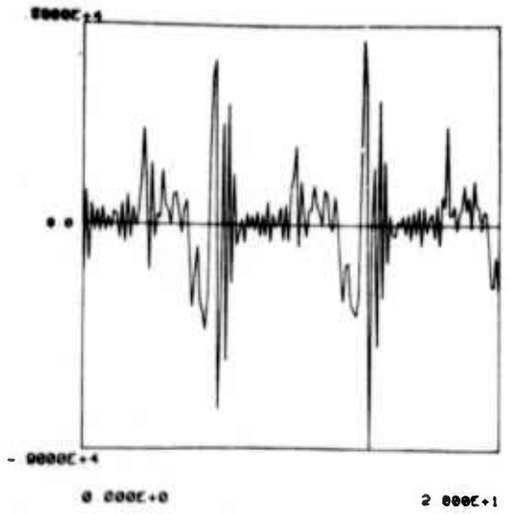
In approximating the error signal as

$$\tilde{e}(n) = \sum_{j=o}^{M} \tilde{b}(j)g(n-j) \quad ,$$

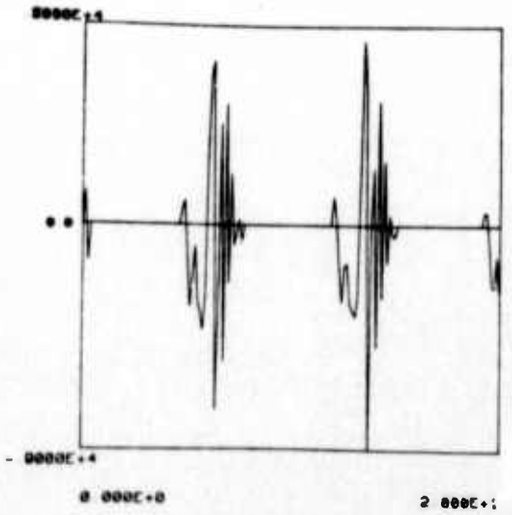
it was stated that the zeros will have no direct effect when g(n) = 0. Since the glottal excitation is also often zero in actual speech production, one of the goals of this research was to determine whether better estimates of the pole coefficients could be obtained by analyzing voiced speech only when the glottis was closed, the speech waveform primarily representing the vocal tract during that state. A decrease in computational effort for calculating the covariance matrix would also be achieved. Prior to loading the covariance matrix to calculate the a(i) coefficients, weighted least squares was used to zero the sections of a pitch period corresponding to the glottis-open state. Speech occurring during the glottis-closed time was weighted by one. The glottis-open segment of a pitch period was set as a fixed percentage of the pitch period. Closure of the glottis was set at the maximum absolute excursion of the waveform in the next pitch period. Results indicated that while use of only 50% of the data in a pitch period might produce a satisfactory synthesis for one phoneme, the



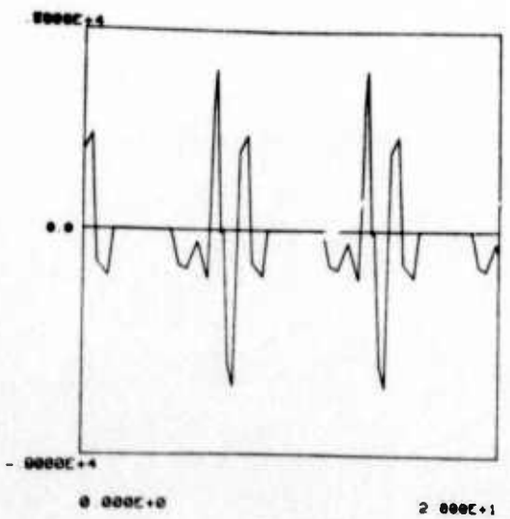
(a)



(b)



(c)



(d)

Figure 2: Example waveforms for the vowel /æ/.
a) Original speech waveform.
b) The error signal e(n).
c) e(r) windowed to retain region near open glottis.
d) Approximate error signal generated from 7 zero coefficients.

synthesis for another phoneme might be completely unacceptable. For this reason all of the speech data is used in computing the covariance matrix.

Synthetic speech generated using the glottal waveform modelling technique described here is intelligible. However, the speech has a muffled sound and lacks the "sharpness" of the original or all-pole LPC synthesis. This muffled quality results from a lack of high frequency energy in the excitation signal generated from the zero coefficients and glottal waveform model. Presently, work is being done on improving the synthesis by eliminating the muffled effect. This seems to require modifications in the model -- especially the derivation of an excitation signal for the all-pole portion of the synthesizer.

REFERENCES

[1] Atal, B.S. and S.L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", Journal of the Acoustical Society of America, Vol. 50 (1971), pp. 637-655.

[2] Rosenberg, A.E., "Effect of Glottal Pulse Shape on the Quality of Natural Vowels", Journal of the Acoustical Society of America, Vol. 49 (1971), pp. 583-590.

[3] Holmes, John N., "The Influence of Glottal Waveform on the Naturalness of Speech from a Parallel Formant Synthesizer", IEEE Transactions on Audio and Electroacoustics, Vol. AU-21, No. 3 (June 1973), pp. 298-305.

SECTION 7

PERCEPTUALLY INVARIANT TRANSFORM ANALYSIS

James T. Kajiya

The research proposed here has as its roots the work of Stockham[1]. There the problem of processing images was approached with a perceptual model in hand. This approach differs radically from conventional methods in that it does not use as its principal model one that describes the method of image production. Other methods of sensory information processing such as Linear Predictive Coding, Blind Deconvolution, etc. attempt to find parameters that describe the functioning of the production mechanism. In [1] Stockham deals with the mechanism that consumes the sensory information.

Image understanding and speech understanding may be approached in this way. The model for the perceptual mechanism however must be chosen in a new way. Fortunately some work has already occured toward this goal.

In 1947 Pitts and McCulloch[2] recognized that images and sounds can be represented as real-valued functions on what they termed the visual and aural manifold. Along with various physiological speculations they recognized that there exist transformations on the manifold that are useful in analyzing the process of perception. This set of

transformations form a Group acting on what can be considered a homogeneous space. Pitts and McCulloch's fundamental observation was that the group of transformations must leave the so called preceptual constancies invariant.

W.C. Hoffman[3,4,5] developed the group of visual transformations and recognized that it was endowed with a Lie Group structure. Basing his work on the experiments of Hubel and Wiesel he was eventually able to predict some optical illusions of angle.

Analysis of functions defined on a homogeneous space also has a rich tradition in modern mathematics and physics.

Powerful methods in Quantum Mechanics used to solve the Schrodinger eigenvalue problem capitalize upon symmetries in the Hamiltonian (see Weyl[6]). These symmetries are expressed as invariances under certain transformations of spacetime,viz. typically the group of three-dimensional rotations about the nucleus (O(3)) or the Lorentz group. Using the theory of Group Representations these methods decompose the space of complex-valued functions defined on the spacetime manifold into invariant irreducible subspaces. These invariant irreducible subspaces for the group O(3) are finite dimensional and are spanned by the so called Spherical Harmonics. The situation for the Lorentz group (see Wigner[7]) is not quite so felicitous. Its corresponding invariant irreducible subspaces need not be

finite dimensional. This fact explains in part a major obstacle to a satisfactory theory of Relativistic Quantum Mechanics. To gain an insight into the difference between the two cases cited above we may ask why they are different. The crucial property we seek is that of the topological notion of compactness. The group O(3) is compact while the Lorentz group is noncompact. It is known that all Lie Groups enjoy the relaxed condition of local compactness.

In the early 1950's G.W.Mackey[8,9,10,11] began to successfully address the issues of infinite dimensional locally compact group representations by unitary operators on a hilbert space. His work made a significant contribution to the progress of the theory. Particularly fundamental was his refinement and extension of the concept of induced representations first used by Frobenius.

An analysis of some of the more basic perceptual Lie Transformation Groups shows that they satisfy the algebraic property of Solvability. The analysis of unitary representations of solvable Lie Groups quite recently initiated by Kirillov, Diximier, Auslander, Moore, et.al.[12,13,14,15,16,17] can thus be inferred to have a major impact on the theory of perception. In fact, it is easy to see that the situation parallels closely that of Quantum Mechanics. Indeed, the wave function is a complex-valued function on the spacetime manifold; similarly images and sounds are real-valued functions on the visual

and aural manifold. Instead of analysis of a probability wave into invariant components we see analysis of images or sounds into components invariant under various perceptual transformations.

Thus by using the theory of Group Representations we hope to obtain a number of transforms whose action expresses closely the invariance that psychophysiologists are wont to call perceptual constancy. Perhaps even more significant is the development of a method that generates perceptually important transforms given models of a perceptual process couched in the language of Lie Groups. It is in this way that we hope to further the cause of information processing in the contex of a psychophysiological model.

REFERENCES

[1] T.G. Stockham,Jr. "Image Processing in the Context of a Visual Model" Proc. IEEE 60(1972) pp.828-842

[2] W. Pitts, W.S. McCulloch "How we know universals: The Perception of Auditory and Visual Forms" Bull. Math. Biophysics 9(1947) pp.127-147

[3] W.C. Hoffman "The Lie Algebra of Visual Perception" J. Math. Psych. 3(1966) pp.65-98

[4] ---------- "Higher visual perception as prolongation of the basic Lie Transformation group" Math. Biosc. 6(1970) pp.437-471

[5] ---------- "Visual illusions of angle as an application of Lie Transformation groups" SIAM Review 13(1971) pp.169-184

[6] H. Weyl THE THEORY OF GROUPS AND QUANTUM MECHANICS New York, 1931

[7] E. Wigner "On unitary representations of the inhomogeneous Lorentz group" Ann. of Math. 40(1939) pp.149-204

[8] G.W. Mackey "Imprimitivity for representations of Locally Compact groups I" Proc. Nat. Acad. Sci. USA 35(1949) pp.537-545

[9] ---------- "Induced representations of Locally compact groups I" Ann. of Math. 55(1952) pp.101-139

[10] ---------- "Induced representations of Locally compact groups II: the Frobenius reciprocity theorem" Ann. of Math. 58(1953) pp.193-221

[11] ---------- "Unitary representations of group extensions I" Acta Math. 99(1958) pp.265-311

[12] A.A. Kirrilov "Unitary Representations of Nilpotent Lie groups" Russ. Math. Surveys 17(1962) pp.53-104

[13] J. Dixmier "Representations induites holomorphes des groupes resolubles algebriques" Bull. Soc. Math. France 94(1966) pp.181-206

[14] L. Auslander, B. Kostant "Polarization and unitary representations of solvable Lie Groups" Invent. Math. 14(1971) pp.255-354

[15] L. Auslander, C.C. Moore "Unitary representations of solvable Lie Groups" Memoirs Amer. Math. Soc. 62(1966)

[16] J. Brezin "Unitary Representation theory for solvable Lie Groups" Memoirs Amer. Math. Soc. 79(1968)

[17] R. L. Lipsman GROUP REPRESENTATIONS Lect. Notes in Math. Springer Verlag (1974)

SECTION 8

Image Understanding

Martin Newell

Work in this new area was conceived and proposed during the early part of this period. Detailed project planning is currently underway.

The analysis of imaging is undertaken for the purposes of automatic recognition of previously known objects, or for synthesizing models of previously unknown objects. This research is based on the hypothesis that such analysis can benefit greatly if carried out in conjunction with three-dimensional models of the objects in the scene.

Given modeling and image synthesis facilities of sufficiently advanced capability, analysis of such scenes can be carried out using an analysis by synthesis approach. The analysis cycle starts with some hypothesis about objects in the actual scene, and their orientation with respect to the camera. A synthetic image is then created with the modeling facility and compared with the actual image. The model of the objects in the synthetic image is then modified based on differences between these two images, and the cycle repeated.

Four main problem areas will be attacked in order to develop such a system.

1. Abstraction of perceptually relevant information from images for use in guiding the comparisons between the real and synthesized images.
2. Techniques for generalized correlation in both two and three dimensions for the purposes of finding the best fit between the real image and a synthetic image.
3. Synthesis of high fidelity images capable of reproducing the perceptually important characteristics of real images.
4. Development of modeling system of sufficiently advanced capability for storing and manipulating a wide variety of object representations.

PUBLICATIONS AND PRESENTATIONS

[1] Baxter, Brent "Image Processing in the Human Visual System" University of Utah, Computer Science Technical Report Number UTEC-CSc-75-168. December 1975.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Date Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER UTEC-CSc-77-017 ✓	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* SENSORY INFORMATION PROCESSING AND SYMBOLIC COMPUTATION		5. TYPE OF REPORT & PERIOD COVERED Semi-Annual 1 July 1975 - 31 Dec. 1975
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s)		8. CONTRACT OR GRANT NUMBER(s) DAHC15-73-C-0363 ✓
9. PERFORMING ORGANIZATION NAME AND ADDRESS Computer Science Department University of Utah Salt Lake City, Utah 84112		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS ARPA Order Number: 2477
11. CONTROLLING OFFICE NAME AND ADDRESS Defense Advanced Research Projects Agency 1400 Wilson Blvd. Arlington, Virginia 22209		12. REPORT DATE January 1976
		13. NUMBER OF PAGES 55
14. MONITORING AGENCY NAME & ADDRESS *(If different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

This document has been approved for public release and sale; its distribution is unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

Same

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

motion blur, color vision, synthetic speech, noise suppression, noise removal, human perception, image understanding, linear prediction, LPC driving function glottal waveform model perceptually, invariant transform analysis.

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

In Section 1, ~~Baxter~~ discusses and demonstrates what parameter adjustments can accomplish in his method of removing motion blur from photographic images.
Section 2 indicates the status of work ~~by Faugeras~~ in color vision, and our expected method of reporting these results.
Section 3. ~~Boll~~ describes a method for strikingly increasing the perceived quality of synthetic speech. Additional computation at the receiver is used to generate two channels (i.e. binaural) of sound for a stereo headphone set. This method requires no change in the existing generation and transmission processes

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

SECURITY CLASSIFICATION OF THIS PAGE *(When Date Entered)*

20 Abstract continuation

and algorithms.

In Section 4. Petersen demonstrates one method of removing noise from speech. Intelligible speech has been generated from input signals that contain +18dB of noise.

In Section 5. Callahn reviews his method of suppressing noise in a one-dimensional signal stream (e.g. speech) by using two-dimensional processing. A complete Technical Report UTEC-CSc-76-209 will be available approximately concurrently with this report.

Sections 6. and 7. report the start of work in the coding of speech, and in the mathematical theory of human perception. Both indicate our future directions in these fields.

Section 8. outlines the Image Understanding Research by Newell that has been proposed for the future.